The Welch Medical Library Indexing Project*

By Sanford V. Larkey, M.D.

Director and Librarian, Welch Medical Library
Johns Hopkins University, Baltimore, Md.

I AM very happy to have this opportunity to report to the Association on some of the aspects of the Welch Medical Library Indexing Project, which has been going on now for about three and one-half years. This research program is supported by the Armed Forces Medical Library through a contract between it and the Johns Hopkins University.

The background of the project—how the Welch Library, after World War II, was interested in a program for training medical librarians with the necessary research aspects; the interest expressed by the then Army Medical Library in the research side of the program; and the eventual negotiation of a research contract—has already been given in the Bulletin in the report of the meeting of the Honorary Consultants to the Army Medical Library of 1948[1]; so I will not repeat the details here.

Our own plans covered a fairly large area of medical bibliography, but with particular emphasis on the problems of indexing of periodical literature, including, by extension, the related aspects of abstract journals, review journals, etc. I believe the statement of the scope of the project, as given in the contract, will give some idea of the general picture:

"To study the problems of indexing medical literature.

To explore the theory and practice of subject heading (nomenclature) and classification (coding) as they concern medical literature.

To explore existing and projected methods, emphasizing machine methods, applicable to medical bibliography, operating such pilot projects as may be necessary, and to report on the suitability of machine methods in the bibliographic operations of the Army Medical Library."

As you can see, the first sentence covers a very large area and really states the whole problem. When taken in connection with the other two points, this charge has worked out in practice to mean the evaluation and study of present

* Read at the 51st Annual Meeting, Medical Library Association, Lake Placid, N. Y., June 24–27, 1952.

[1] Larkey, Sanford V. The Army Medical Library Research Project at the Welch Medical Library. Bulletin, 37: 121–124, April 1949.

indexing and abstracting services, including studies of their coverage of the medical literature, of their methods of indexing, and of their use.

One way to learn something about how indexing and abstracting services are used is to ask those who use them. We decided to do this, asking directly through interviews rather than by questionnaires. We interviewed a fairly large number of medical librarians and scientists, and many of you generously took part in such interviews at the Galveston meeting. The analysis and study of these interviews gave us many valuable leads, but in the meantime there have been so many changes in the indexes and abstract journals, that further studies along these lines will be needed.

Studies have been made of the coverage of journals and of articles within journals by individual services, but a detailed analysis of the coverage of all or almost all medical literature requires a knowledge of what periodicals are available in the world in medicine and related fields. Since a principal objective of the project in this regard was to determine the scope and coverage desirable in a medical index, with special reference to one published by the Armed Forces Medical Library, it was decided to compile a list of the world's medical journals, on IBM punched cards, with facts included which would give information as to scope, coverage by indexing and abstracting services, and features for evaluation of the journals as to their possible inclusion in a medical index. Details of the present status of this part of the project will be given later in the paper.

Similarly, I will defer discussion of the other aspects, namely subject indexing and subject headings, and machine methods, although work on all phases has been proceeding simultaneously. Since progress reports on all aspects of the project have been given every year to the Honorary Consultants to the Army Medical Library and published in the BULLETIN[2], it is not intended to review these developments but to tell now only of the present status of certain things we are doing.

PSYCHOLOGY OF MACHINES

I think I should say something about "machines" themselves at this point. Since we are using machines in all the major phases of our work, I would like to describe the machines we are using and just how we are using them. I will discuss the present status of each phase of our work primarily on the basis of the machine operations involved. Another reason for this approach is that we have found in discussing our program with others, our use of machines seems either to interest or worry people more than any other feature.

This brings me to what might be called the "psychology of machines." The very word "machines" seems to do things to people. We hear talk of "electronic robots," as though they were some sort of "men from Mars" who could take

[2] BULLETIN, 38: 113–116, April 1950; 39: 87–89, April 1951; 40: 107–112, April 1952.

over all intellectual activities by merely pushing buttons. This sort of talk leads to excessive hopes or to inordinate fears and precludes objective thinking about the possible uses of machines. One should consider machines as practical adjuncts, as we do typewriters, 3 x 5 cards, and visible indexes. Machines are only doing very rapidly what one could do with his own eyes and brain if he had all the time in the world to do it and wanted to do it. There is no magic about it.

There is, however, a more valid psychological aspect to machines. Since machines operate on a strict yes-or-no principle, we must be rigidly exact in presenting a problem. Each step must be in the most precise logical form, since one rarely can stop to correct as one goes along. Each step must be gone over and over in relation to every other one. One has to think not once, but many times. Programming often takes almost as long as the machine operation itself, but the end result is still reached much more quickly than by manual operations.

These strict limitations of machines have been very useful to us. They not only have tightened up our own thinking processes, but their application has emphasized many semantic inconsistencies in our terminology and classifications. So, perhaps there may be a good psychological side to machines.

Machines and Our Use of Them

Our machines at least have a very respectable genealogy. As Raymond Pearl[3] has pointed out, it was John Shaw Billings, who, while working on the 1880 census, suggested to Mr. Herman Hollerith the idea of punching holes in cards for recording data and then sorting by mechanical means. Mr. Hollerith invented the basic IBM machines. It is not known if Billings ever thought of applying the principle to bibliographical work, but it would seem eminently fitting that it might be so utilized.

The machines we are now using in the project are all IBM equipment and I will confine myself largely to a description of these and of our uses of them. Most of these machines were designed primarily for statistical and accounting uses, including tabulating, and their application to or modification for our purposes is one of the tasks confronting us. There are, however, many operations which are applicable to our purposes, and we are trying to make the best use of them.

I think it will be well to describe what the machines can do in some detail, since it will make clearer what I will have to say about our application of them. I will speak of these functions in terms of the card, because cards are our units of information, on which have been punched items of information, either numerically or alphabetically as words or abbreviations. Codes could be either or both. Probably the basic thing a machine can do is to *sort*, that is, pick out

[3] Pearl, Raymond. Some contributions of Dr. John Shaw Billings to the development of vital statistics. Bull. Inst. Hist. Med. 6: 387–393, May 1938.

cards which have a common characteristic or characteristics and put them together in packs. A machine can *count* these cards or certain points of information punched in the cards. Another machine, the collator, can *match* like cards or various punches in cards, and *merge* two decks of cards in a numerical or alphabetical sequence. Other machines, the tabulators and card-operated typewriters, can *print* what has been punched on cards.

A special type of sorting is what has been termed "searching." This involves the selecting out of cards which meet complex requirements of punches or combinations of punches. This procedure can be carried out with some of the ordinary standard machines, but would take so many steps that it would not be practical in point of view of time and effort except for very small files of cards. The problem is now being approached by means of much more elaborate IBM machines—the 101 and another designed especially for this purpose.

We are using or have some familiarity with all of these IBM machines. The first group of machines are those used for preparing the cards. Our cards are prepared with an interpreting punch which simultaneously prints on the card what has been punched. Cards are then checked on the verifier if necessary. A reproducer is used when duplication of a large number of cards is desired or when there is a common item to be punched into a number of cards. This latter operation is known as "gang punching." By use of prepared number decks or special attachments, continuous serial numbers can be punched in a file of cards. We have used other standard machines for sorting, collating, and printing. We are now using the IBM 101 electronic statistical machine for most of our sorting but particularly for searching.

IBM machines are activated by an electrical current set up by contact of brushes through the holes punched in the cards. The operation of the ordinary sorter is relatively simple. Operating on only one column at a time, an electric impulse is initiated by the brush contact through the hole in the card. The impulse travels through a set of electro-magnetic relays and causes a chute to open. This directs the card to a pocket corresponding to the punch in the column. But it is when more complex patterns are needed that the electronic element comes in, as in the 101; for here, a great number of matching operations, that is, matches between what is desired and what is punched in a card, are performed in an amazingly short period of time. This is accomplished by setting up on a wiring board, somewhat like a telephone switchboard, extensive and complicated selector circuits which send currents through series of electronically controlled relays. Such electronic set-ups are used in many of the more complicated IBM machines, and notably in the 101. Such a system is almost essential for any high speed searching device. Our work covers a wide range of these operations, from the simplest to some of the most complicated. In some phases we have used machines almost entirely as a research tool while in others the machine is the all-important or essential feature. I intend now to

describe certain areas of our work, emphasizing the role of the machines, but also to tell you something of the overall status of our work in these areas.

Subject Heading Studies

One of the great difficulties in compiling and in using indexes of any kind has been that of the subject headings themselves, the essential key to the index. There has been little standardization of headings used by various indexes and even inconsistencies in usage within a single index. A great deal of this trouble stems from basic inadequacies and inconsistencies in medical terminology and nomenclature. Subject headings have attempted to keep up with changing terminology, but since the headings are usually in alphabetical lists, it is very difficult to work out relationships. It was our plan to transform alphabetical lists of subject headings to a categorized arrangement. We started by using conventional 3 x 5 cards, typing on these cards the headings and cross references with complete tracings, from *Quarterly Cumulative Index Medicus*, *Index-Catalogue*, and other sources. Category arrangements were then attempted. Preliminary results of these studies convinced us that we were on the right track. For instance, striking discrepancies were shown up in the field of chemistry.

It was not until last year that we began to use punched cards for this part of our work. We started then to put on punched cards all the headings used by *Current List* in the first five months of 1951. The actual heading was put down in full by alphabetical punching and code numbers punched for category sorting. As it turned out, we were very glad we had gone over to punched cards, for this file was the basis for the later studies for revision of the *Current List* headings now being used.

I am not going to say much about the basis for the revision since Mr. Seymour Taine will tell of that. I will only mention some of the machine features. By category sorting and then printing lists from the punched cards, we were able to study the terms used in restricted fields and to compare them against other authorities. For instance, having based our category code for diseases on the World Health Organization International Statistical Classification, we sorted on these code numbers and then could compare the terms used by *Current List* with all the terms in the WHO classification for each specific group of diseases. The staff of the *Current List* and the project cooperated on similar studies for all of the categories, and the revised cards constituted the main subject heading deck.

See and *See also* references were punched on cards and by machine methods put in the proper order with the main subject headings. Lists were then printed from the cards. This was the tentative authority list. The next step was to add the tracings, *See from* and *See also from* cards. A good part of the compilation and punching of these was done by machine methods. The last printing was

done directly onto Multilith mats, making it possible to produce a large number of copies of the list. Here we have made most extensive utilization of the *printing* function of the machines.

We will continue our category studies, with comparison of terms from many other sources. While we realize we can never have a permanent standard authority, we hope we can make an approach to something like this and to have a method for logical and fairly easy revision.

Journal List on Punched Cards

In the compilation of the list of the world's medical serials on punched cards, the machines were intended to be used primarily as research tools. Since analyses of scope, coverage by indexing and abstracting services, and evaluation of the thousands of titles require compilation and study from a number of different axes, it was believed that machines would be a great saving of time and would provide more extensive and accurate statistical data. At the same time, the sorting and printing features of the machines would enable us to have preliminary groupings or listings along any line desired. We now have about 7000 serial titles represented in our file.

The information about each journal has been taken from the journal itself, from the latest issue for current information (title, sponsorship, country, publisher, language, frequency, etc.) and from a whole volume for more general information (type, contents, and subject fields covered). To date, information has been assembled from the Armed Forces Medical Library, the Library of the American Medical Association, and the Welch Library. Preliminary checks have been made against the holdings of the New York Academy of Medicine and the Library of National Institutes of Health. The staff of the Armed Forces Medical Library is supplying us with data sheets of information on its holdings and we could not have done anything without its help.

After the information is collected, it is then punched on cards as follows: a number code for alphabetizing, abbreviation of the title, code numbers for language, country, frequency, type, contents, and major and minor subject fields. In addition to this information about the journal itself, other types of information are being assembled and punched in code on the cards. These include coverage by indexing and abstracting services, holdings by libraries, inclusion on selective lists, such as the MLA list of recommended periodicals.

We are thus able to make analyses from many different points of view. We have made a number of preliminary studies which have included statistical studies of coverage by *Current List*, *QCIM*, *Excerpta Medica*, *Biological Abstracts*, and *Chemical Abstracts*, comparative studies of the coverage by combinations of these, studies of coverage by country and by subject by *Current List* and by *QCIM*, and listing of indexing and abstracting services, and of other journals publishing indexes and/or abstracts. Some of the studies

have been previously reported in the BULLETIN. I might say that one form of information most difficult to come by is that of coverage. Very few services publish lists of the journals they cover and we have been eagerly awaiting the publication of *Periodica Medica Mundi*. If it does not appear soon, we may have to go through the entire year's file of a service to pick out the journals covered. Even this might be worthwhile, for we believe an analysis of this coverage would be a valuable thing in itself and a factor in evaluation.

Our earlier analyses were done with a standard Sorter. Now that we are using the 101 machine, we will be able to make more complicated analyses, generally at one pass, and also have counts on all items registered by machine. At the present time we are working at bringing our file up to date, planning for adding information on additional items of information that we can dig out. We plan to set a deadline in the late fall and then to begin a series of detailed analyses, leading to evaluation studies.

One of the advantages of having such a file on punched cards is that it can easily be integrated with similar punched card files, such as those proposed by the Library of Congress. At present we have a limitation in the printing feature of the machine, in that we can print from our card only an abbreviated title. This is really all that is necessary for a research tool, but as the demand arises for printed lists with full title and publisher, we will have to provide additional cards to do this. We have worked out designs for these.

MACHINE INDEXING AND SEARCHING

The term machine indexing has acquired two rather different meanings. In the first place it may mean the preparing and printing of a printed index by machine methods, or it may mean an information searching system wherein the various subject concepts of a document are represented in code on punched cards, or on other media, and a machine is used to select out the desired items.

There are a number of instances of printed indexes produced by machines but not, so far as I know, in the field of medicine. We are considering, on an experimental basis, the possibilities of preparing the entire *Current List*, by machine methods. Mr. Eugene Garfield of the project has worked out a detailed procedure, utilizing a number of IBM machines. I want to emphasize that the machine cannot perform the intellectual side of reading articles and of subject indexing. The machines take over only after this has been done. All of the entries, as they now appear in the *Current List*, in the Register and in the Subject and Author indexes, will be on punched cards. All of the numbering, sorting, merging with subject and cross reference entries, and arranging in alphabetical and subject order will be done by machine. The copy for photo-offset printing, except at present for the Register entry, will be printed from the punches in the cards. I will not go into any details now, as we are just beginning our first pilot run. If the method appears to be practicable, we plan to study its application to other types and formats of indexes.

Because there has been a feeling that modern science demands a more rapid, a more thorough, and a more detailed correlation of the facts set forth in scientific literature than apparently is possible with present printed or card indexes, attention has turned to machines as a way of accomplishing this. In a machine searching operation, the facts or subject concepts of a document must be symbolized by codes which can then be represented by punches in a card. The fundamental practices of analysing a document, let us say a periodical article, are essentially the same as in present subject indexing. One must still read the article to find what it is about and what are its salient facts. Assuming this, the various subject concepts in the article are assigned code designations from a pre-arranged code. The code designations for all of the subject concepts in an article are then punched on a card or cards, along with a serial number identifying the document. It is the code numbers for which the machine searches, correlating those in the card or cards representing a single document for answers to a specific question requiring a certain combination of code numbers.

Suppose, for a simple example, the number 90015 represents the concept *pulmonary tuberculosis* and 35062 *Streptomycin.* If one wanted all articles on streptomycin treatment of pulmonary tuberculosis, the machine would be set to pick out all those cards which have both of these numbers punched in them. In our earlier work with a sorter and collator, it was necessary to have the code numbers for various categories in a fixed order and in a fixed place on the card. This was also true of earlier work with the 101 machine, but now Mr. Garfield has worked out an amazing but fundamental wiring system for the 101 which permits searching for any 5-digit number in any one of 16 five-column fields on a card. His system permits searching for up to 48 such 5 digit numbers in various combinations at one time. This will allow a number of separate questions to be asked at one time.

At present we are experimenting with arbitrary numerical codes but will soon use actual coding from indexing articles. I will not give any details since the project will prepare a comprehensive report. One should realize though what the machine is doing here. As described earlier, there is a very rapid matching operation going on for each card, correlating the desired combinations of all these 48 individual codes with those codes punched in the card. The cards are going through the machine at the rate of 450 a minute.

Machines may speed up searching operations; they should permit greater depth of indexing and possibly greater coverage of the literature; and there should be greater possibilities of correlation since the machine can search for any combination of concepts. Whether these advantages will outweigh the values of printed indexes or partially so, only time and more study will tell.

Summary and Conclusion

I have tried to tell you of some of the problems we have been working on and where they are leading us. I still believe, as I said almost four years ago, before

the project started, "that our present indexes are extremely valuable bibliographical aids and that while a search is being made for new methods, including machine methods, the improvement of these indexes in something like their present form must be a primary consideration."[4] A great deal of our work has been along these lines. It is quite possible that machines will help in solving some of the problems. We have found that their practical application has been greater and more useful than we had anticipated. Familiarity has bred anything but contempt.

In conclusion, I should like to point out that, while our work and our findings are primarily for the use of the Armed Forces Medical Library, we feel that they will be applicable to other fields just as we have profited greatly from work others are doing. Coordination of research in bibliographical problems in all fields should benefit all.

[4] Larkey, S. V., op. cit.